

Data Engineering Project

Tokyo Olympic - AZURE).

Réalisé par :

Boujbair Oussamae

Supervisé par :

MOI MM

DEMOS

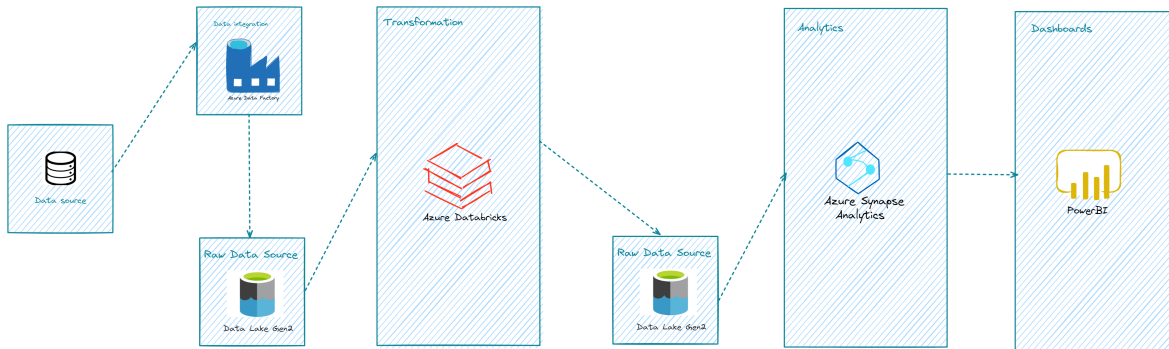
07 DEC 2023
Rabat, Maroc

Contents

1	Project architecture	2
2	Data Source (GitHub Repo)	2
2.1	GitHub: 5 files	2
3	Azure	3
3.1	Création compte de stockage	3
3.2	Création conteneur:	3
3.3	Data Factory	4
3.4	Azure Databricks	4
3.5	Azure synapse analytics	5

1 Project architecture

Voici l'architecture du projet:



2 Data Source (GitHub Repo)

2.1 GitHub: 5 files

On a créé un repo sur github et on a chargé 5 fichiers qu'on va exploiter par la suite:

Azure-DataEngineering / Data /

Add file

Boujbair

Delete Data/ok

27c9bc3 · yesterday

History

Name	Last commit message	Last commit date
..		
Athletes.csv	Add files via upload	yesterday
Coaches.csv	Add files via upload	yesterday
EntriesGender.csv	Add files via upload	yesterday
Medals.csv	Add files via upload	yesterday
Teams.csv	Add files via upload	yesterday

3 Azure

Services Azure

- Créer une ressource
- Comptes de stockage
- Inscriptions d'applications
- Azure Databricks
- Machines virtuelles
- Abonnements
- Centre de démarrage...
- App Services
- Bases de données SQL
- Autres services

Ressources

Récent Favori

Nom	Type	Dernier affichage
tokyoolympicdataoussama	Compte de stockage	il y a quelques secondes
tokyo-olympic-sa-ouss	Espace de travail Synapse	il y a 17 heures
Azure-data-eng	Groupe de ressources	il y a 17 heures
tokyo-olympic-dfouss	Fabrique de données (V2)	il y a un jour

Tout afficher

Naviguer

- Abonnements
- Groupes de ressources
- Toutes les ressources
- Tableau de bord

3.1 Création compte de stockage

tokyoolympicdataoussama
Compte de stockage

Rechercher

Vue d'ensemble

- Journal d'activité
- Étiquettes
- Diagnostic et résoudre les problèmes
- Contrôle d'accès (IAM)
- Migration des données
- Événements
- Navigateur de stockage
- Stockage des données
- Conteneurs
- Partages de fichiers
- Files d'attente
- Tables
- Sécurité + réseau
- Mise en réseau

Bases

Groupe de ressources (déplacer)
Azure-data-eng

Emplacement
southafricanorth

Emplacement principal/secondaire
Principal : South Africa North, secondaire : South Africa West

Abonnement (déplacer)
Azure subscription_1

ID d'abonnement
a526b245-03fc-4e14-89b0-127e3954bf37

État du disque
Principal : Disponible, secondaire : Disponible

Étiquettes (modifier)
Ajouter des étiquettes

Propriétés

Supervision Fonctionnalités (5) Recommandations (0) Tutoriels

Outils et Kits de développement logiciel (SDK) de développement logiciel (SDK)

3.2 Création conteneur:

On crée un conteneur et on crée deux fichiers raw-data et transform data dans ce conteneur.

tokyoolympicdata
Conteneur

Rechercher

Vue d'ensemble

- Diagnostic et résoudre les problèmes
- Contrôle d'accès (IAM)
- Paramètres
- Jetons d'accès partagé
- Gérer l'ACL
- Stratégie d'accès
- Propriétés
- Métadonnées

Bases

Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)
Emplacement : tokyoolympicdata

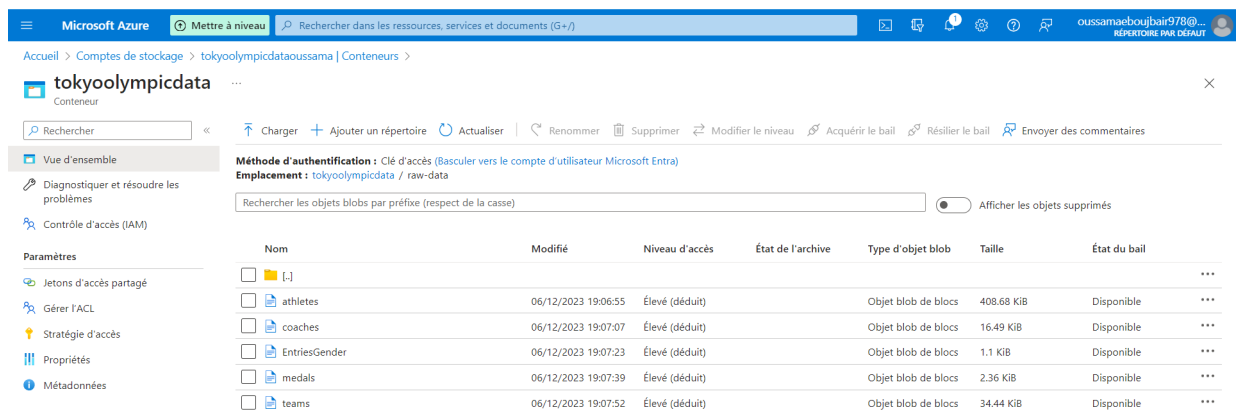
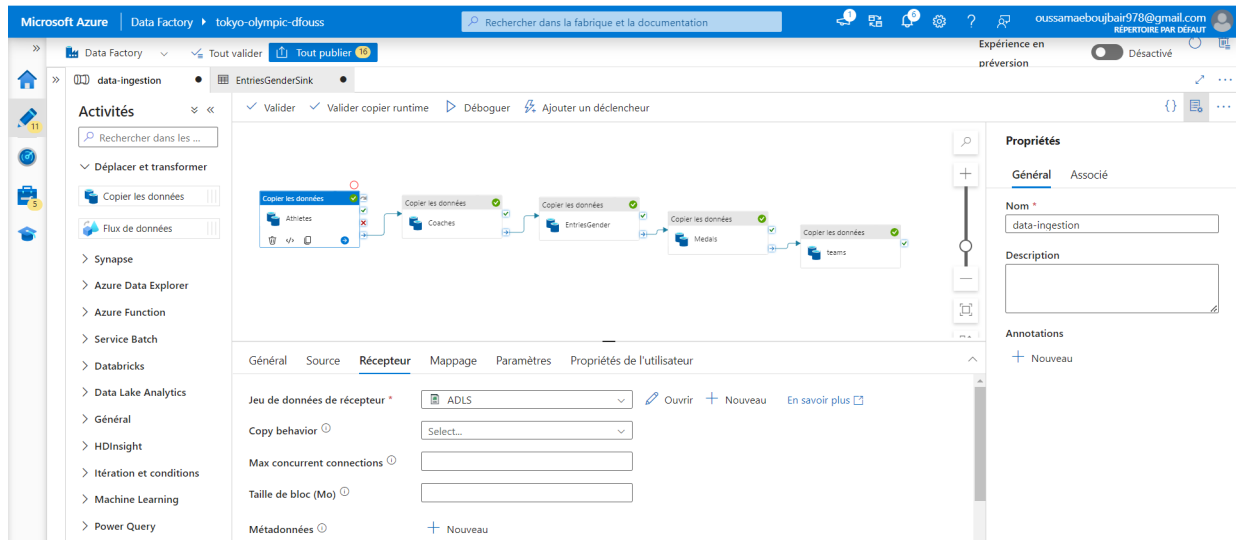
Rechercher les objets blobs par préfixe (respect de la casse)

Afficher les objets supprimés

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bail
raw-data						...
transform-data						...

3.3 Data Factory

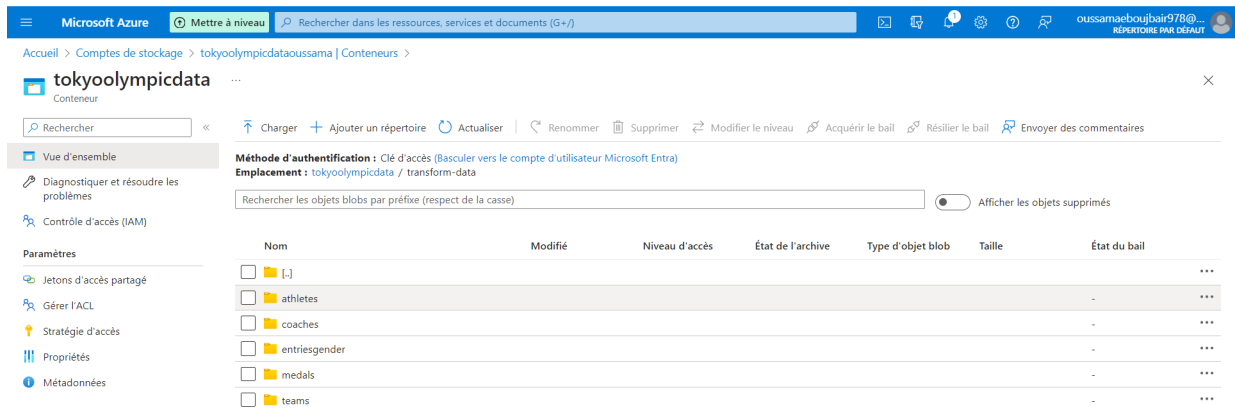
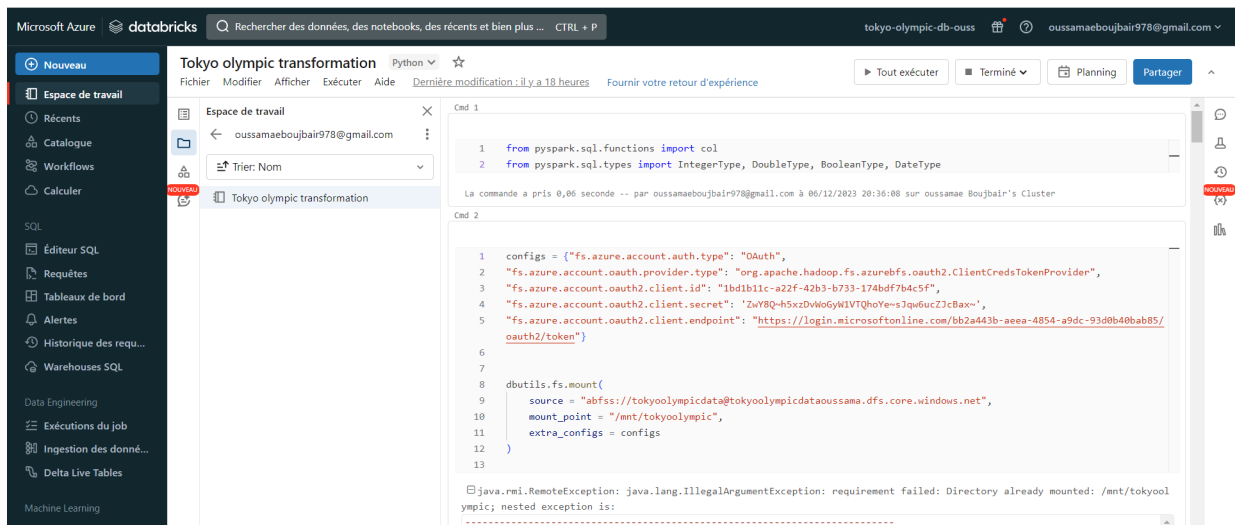
On importe les données de la source (github repo) dans data lake gen 2 (raw-data), on utilisant lien HTTP. Data factory permet d'ingester les données from diff sources, dans notre cas c'est github repo.



3.4 Azure Databricks

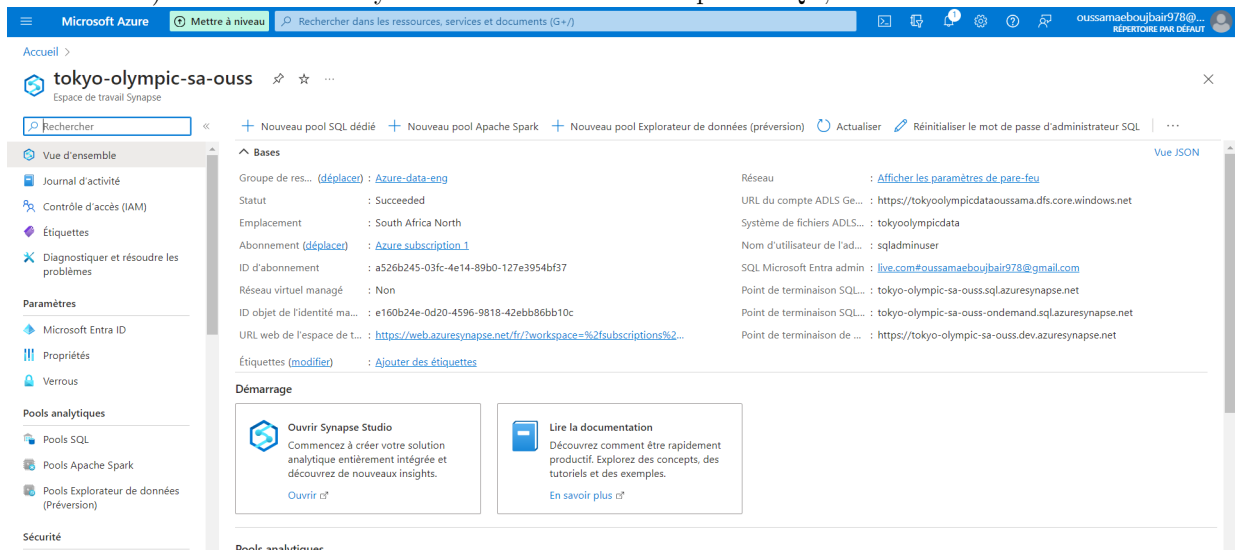
On crée un noutebook sur Databricks pour transformer les données.

On import les données de data lake gen 2 (Raw-data) avec script pyspark et on fait les transformations neccessaires et on charge les données transformées dans data lake gen2 (transform data)



3.5 Azure synapse analytics

On crée un compte azure synapse analytics et on importe les données transformées from data lake gen 2 (transform data) et on fais les analyses necessaires avec des requetes SQL, et on des visualisations.

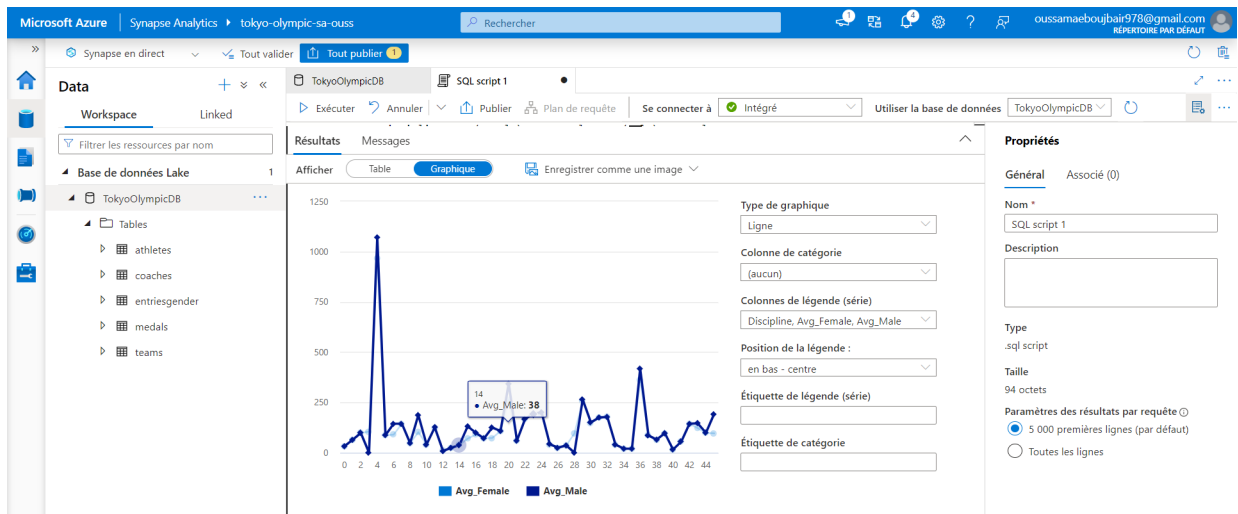


The screenshot shows the Microsoft Azure Synapse Analytics interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and 'tokyo-olympic-sa-ouss'. The left sidebar shows the 'Data' section with 'Workspace' and 'Linked' options. The 'Base de données Lake' section is expanded, showing 'TokyoOlympicDB' and its tables: 'athletes', 'coaches', 'entriesgender', 'medals', and 'teams'. The 'SQL script 1' tab is active, displaying the following SQL query:

```
1 SELECT Discipline, AVG(Female) Avg_Female, AVG(Male) Avg_Male
2 FROM entriesgender
3 Group by Discipline;
```

The 'Propriétés' panel on the right shows the 'Général' tab with the following details:

- Nom *: SQL script 1
- Description: (empty)
- Type: .sql script
- Taille: 94 octets
- Paramètres des résultats par requête: 5 000 premières lignes (par défaut)



FIN